Characterizing the Performance of the Conway-Maxwell Poisson Generalized Linear Model

Royce A. Francis[1,2], Srinivas Reddy Geedipally[3], Seth D. Guikema[2], Soma Sekhar Dhavala[5], Dominique Lord[4], Sarah LaRocca[2]

[1]Department of Engineering Management and Systems Engineering, George Washington University, 1776 G St., NW #159, Washington, DC 20059
{seed@gwu.edu}

[2]Department of Geography and Environmental Engineering, Johns Hopkins University, 313 Ames Hall, 3400 N. Charles St., Baltimore, MD 21218
{sguikema@jhu.edu, larocca@jhu.edu}

[3]Texas Transportation Institute, Texas A&M University, 3135 TAMU, College Station, TX 77843-3135
{srinivas8@tamu.edu}

[4]Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136
{d-lord@tamu.edu}

[5]Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143
{soma@stat.tamu.edu}

Corresponding Author Information: Royce A. Francis, Department of Engineering Management and Systems Engineering, George Washington University, 1776 G St., NW #159, Washington, DC 20059. E-mail: seed@gwu.edu; Phone: 412-589-9865.

Authors' Footnote: Royce Francis is Postdoctoral Fellow, Seth Guikema is Assistant Professor, and Sarah LaRocca is Graduate Research Assistant in the Department of Geography and Environmental Engineering, The Johns Hopkins University, Baltimore, MD 21218 (e-mail: royce.francis@jhu.edu, sguikema@jhu.edu, larocca@jhu.edu). Srinivas Geedipally is Engineering Research Associate, Dominique Lord is Assistant Professor, Texas Transportation Institute, College Station, TX 77843 (e-mail: srinivas8@tamu.edu, d-lord@tamu.edu). Soma Dhavala is PhD Candidate, Department of Statistics, Texas A&M University, College Station, TX 77843 (email: soma@stat.tamu.edu). This work was partially supported by the National Science Foundation (grant ECCS-0725823), the U.S. Department of Energy (grant BER- FG02-08ER64644) and the Whiting School of Engineering. All opinions in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

Running Head: Characterizing the COM-Poisson GLM

ABSTRACT

Count data is pervasive in many areas of risk analysis; deaths, adverse health outcomes, infrastructure system failures, and traffic accidents are all recorded as count events for example. Risk analysts often wish to estimate the probability distribution for the number of discrete events as part of doing a risk assessment. Traditional count data regression models of the type often used in risk assessment for this problem suffer from limitations due to the assumed variance structure. A more flexible model based on the Conway-Maxwell Poisson (COM-Poisson) distribution was recently proposed, a model that has the potential to overcome the limitations of the traditional model. However, the statistical performance of this new model has not yet been fully characterized. This paper assesses the performance of a maximum likelihood estimation method for fitting the COM-Poisson generalized linear model (GLM). The objectives of this paper are to (1) characterize the parameter estimation accuracy of the MLE implementation of the COM-Poisson GLM and (2) estimate the prediction accuracy of the COM-Poisson GLM using simulated datasets. The results of the study indicate that the COM-Poisson GLM is flexible enough to model under-, equi- and overdispersed datasets with different sample mean values. The results also show that the COM-Poisson GLM yields accurate parameter estimates. The COM-Poisson GLM provides a promising and flexible approach for performing count data regression.

Keywords: underdispersed count data, overdispersed count data, maximum likelihood estimation, regression

1. INTRODUCTION

Assessing risk on the basis of count data is important in many areas of risk analysis, including modeling disease prevalence as a function of exposure and other explanatory factors, assessing accident risk in transportation networks, and modeling infrastructure performance during disasters. Generalized linear models (GLMs) are often used to estimate the conditional probability mass function (PMF) for the count event *y* given explanatory information contained in the matrix ***x***. Denoting this PMF as $P(y|\mathbf{x})$, a GLM consists of a conditional mass function and a function linking the information in ***x*** to one or more parameters of the conditional distribution for *y*. For example, a Poisson GLM, a common GLM used for count data, is given by

$$P(y \mid x) \sim Poisson(\lambda) \tag{1}$$

$$\log(\lambda) = \beta_0 + \sum\nolimits_j \beta_j x \tag{2}$$

where the parameters in the vector β are estimated based on the data. Other distributions beyond the Poisson distribution can be used for the conditional mass function (e.g., Poisson-gamma or Poisson-lognormal), and we focus in this paper on the performance of a GLM in which the Poisson distribution is replaced by the COM-Poisson distribution as discussed below.

Count data GLMs are widely used in risk analysis. For example, Liu et al.(1), Han et al.(2, 3) and Guikema and Coffelt (4) use GLMs and Generalized Additive Models (GAMS, a semi-parametric extension of GLMs) to estimate the number and location of power outages and damage to power distribution systems during hurricanes. Maher and Summersgill(5), Lord et al.(6, 7), and Anastasopolous et al.(8) use GLMs to estimate the number of traffic accidents; see Lord and Mannering(7) for an extensive review of data and modeling issues as well as the latest modeling methods for analyzing crash data. Guikema and Coffelt(4) discuss the importance of count data modeling in risk analysis in general. In all of these examples, the underlying data consists of counts of events – power outages, traffic accidents, and deaths – and the goal is to estimate the PMF, an input of fundamental importance to a comprehensive risk model. The challenge is that traditional GLMs used in these analyses do not provide the flexibility needed to accurately model the variance structure in many real-world data sets (Lord and Mannering(7) as an example).

The Poisson family of discrete distributions stands as a benchmark for analyzing count data. However, because the Poisson distribution itself has well-known limitations due to the assumed mean-variance structure, a number of generalizations have been proposed. The Conway-Maxwell Poisson (COM-Poisson) distribution is one of these generalizations. Originally developed in 1962 as a method for modeling both underdispersed and overdispersed count data(9), the COM-Poisson distribution was then revisited by Shmueli et al.(10) after a period in which it was not widely used. Shmueli et al.(10) derived many of the basic properties of the distribution. The COM-Poisson belongs to the exponential family as well as to the two-parameter power series family of distributions. It introduces an extra parameter, ν, which governs the rate of decay of successive ratios of probabilities. It nests the usual Poisson ($\nu = 1$), geometric ($\nu = 0$) and Bernoulli ($\nu = \infty$) distributions and it allows for both thicker and thinner tails than those of the Poisson distribution(10, 11). The conjugate priors for the parameters of the COM-Poisson distribution have also been derived(12, 13).

The COM-Poisson distribution has recently become much more widely used, including studies analyzing word length(10), birth process models(14), prediction of purchase timing and quantity decisions(11), quarterly sales of clothing(10), internet search engine visits(15), the timing of bid placement and extent of multiple bidding(12), developing cure rate survival models(16), modeling electric power system reliability(17), modeling the number of car breakdowns(18), and modeling motor vehicle crashes(6, 19).

Only recently has the COM-Poisson distribution been used in a generalized linear model setting, and the estimation efficiency and bias has not been adequately assessed. The first COM-Poisson GLM was presented at the 2006 Annual Meeting of the *Society for Risk Analysis* and later published in Guikema and Coffelt(17). Lord et al.(6, 19) and Geedipally and Lord(20) then utilized this model to analyze traffic accident data. The approach of Guikema and Coffelt(17) depended on MCMC for fitting a dual-link GLM based on the COM-Poisson distribution starting from non-informative priors on the regression parameters. It also used a reformulation of the COM-Poisson to provide a more direct centering parameter than the original COM-Poisson formulation. Sellers and Shmueli(21) developed an MLE for a single-link GLM based on the original COM-Poisson distribution. We have compared the effects of using these two different link functions and found the estimates of the count events to be nearly identical. Jowaheer and Khan(22) compared the efficiency of quasi-likelihood and MLE estimation approaches for estimating the parameters of a single-link COM-Poisson GLM in the case of equidispersion based on simulated data sets.

Aside from the limited tests of Jowaheer and Khan(22) , there has not been any evaluation of the accuracy or bias of parameter estimates from the COM-Poisson distribution, particularly for cases with underdispersion and overdispersion. Given that a major advantage of the COM-Poisson GLM is its ability to handle both underdispersion and overdispersion within a single conditional distribution, testing the estimation accuracy and bias is a critical need. This paper represents a significant advance in this respect. First, although the MLE for an alternative single-link GLM formulation has been developed by Sellers and Shmueli(21), this paper presents the MLE for the Guikema and Coffelt(17) GLM formulation. Second, although Jowaheer and Khan(22) have compared the Sellers and Shmueli(21) MLE parameter estimation approach to a quasi-likelihood parameter estimation approach, their paper does not demonstrate the accuracy and bias of the GLM. In addition, while Jowaheer and Khan(22) report the standard deviation for parameter estimates as a measure of uncertainty, their investigation seems to imply that these estimates are unbiased. In this paper on the other hand, we comprehensively investigate the accuracy and bias of the COM-GLM, both for prediction and parameter estimation. Moreover, while Jowaheer and Khan(22) find that the COM-GLM quasi-likelihood and MLE approach did not converge in their simulation study in 15% and 45% of cases, respectively, they do not discuss the regions of the sample space for which the COM-GLM model may not converge. In this paper, we discuss potentially problematic regions of the sample space for the estimation of COM-GLM parameters.

The objectives of this paper are to: (1) evaluate the estimation accuracy of the Guikema and Coffelt(17) COM-Poisson GLM for datasets characterized by overdispersion, underdispersion and equidispersion with different means based on a maximum likelihood estimator, and (2) characterize the accuracy of the asymptotic approximation of the mean of the COM-Poisson distribution suggested by Shmueli et al.(10) for the Guikema and Coffelt(17) COM-Poisson GLM. In addition, we briefly discuss potential convergence issues with the infinite series in the normalizing factor of the distribution. We base our analysis on 900 simulated data sets representing 9 different mean-variance relationships spanning underdispersion, overdispersion, and equidispersion for low, moderate, and high means. This more comprehensive characterization of the performance of the MLE-based COM-Poisson regression model provides a strong basis on which to evaluate its usefulness for risk assessment based on count data. This paper is organized as follows. The next section describes the COM-Poisson distribution and its GLM framework. The third section presents our research method. The fourth section gives the results of our computational study. The fifth section gives a brief discussion of the results, and the sixth section provides concluding comments.

2. BACKGROUND

This section describes the characteristics of the COM-Poisson distribution and the COM-Poisson GLM framework.

2.1 Parameterization. The COM-Poisson distribution was first introduced by Conway and Maxwell(9) for modeling queues and service rates. Although the COM-Poisson distribution is not particularly new, it had been largely unstudied and unused until Shmueli et al(10) derived the basic properties of the distribution.

The COM-Poisson distribution is a two-parameter extension of the Poisson distribution that generalizes some well-known distributions including the Poisson, Bernoulli, and geometric distributions(10). It also offers a more flexible alternative to distributions derived from these discrete distributions, such as the binomial and negative binomial distributions. The COM-Poisson distribution can handle both underdispersion (variance less than the mean) and overdispersion (variance greater than the mean). The probability mass function (PMF) of the COM-Poisson for the discrete count Y is given as:

$$P\left(Y = y \middle| \lambda, \nu\right) = \frac{\lambda^y}{\left(y!\right)^\nu} \cdot \left[Z\left(\lambda, \nu\right)\right]^{-1} \tag{3}$$

where:

$$Z\left(\lambda, \nu\right) = \sum_{n=0}^{\infty} \frac{\lambda^n}{\left(n!\right)^\nu} \tag{4}$$
$$y = 0, 1, 2, \ldots, \infty$$
$$\lambda > 0 \text{ and } \nu \geq 0$$

Here λ is a centering parameter that is directly related to the mean of the observations and ν is the shape parameter of the COM-Poisson distribution. Z is the normalizing constant. The condition ν>1 corresponds to underdispersed data, ν<1 to overdispersed data, and ν=1 to equidispersed (Poisson) data. Several common PMFs are special cases of the COM-Poisson with the original formulation. Specifically, setting $\nu = 0$ and $\lambda < 1$ yields the geometric distribution, $\nu \rightarrow \infty$ yields the Bernoulli distribution in the limit, and $\nu = 1$ yields the Poisson distribution. This flexibility greatly expands the types of problems for which the COM-Poisson distribution can be used to model count data.

The exact expressions for the mean and variance of the COM-Poisson derived by Shmueli et al.(10) are given by Equations (5) and (6) below.

$$E\left[Y\right] = \frac{\partial \log Z}{\partial \log \lambda} \tag{5}$$

$$Var\left[Y\right] = \frac{\partial^2 \log Z}{\partial \log^2 \lambda} \tag{6}$$

The COM-Poisson distribution does not have closed-form expressions for its moments in terms of the parameters λ and v. However, the mean can be approximated through a few different approaches, including (i) using the mode, (ii) including only the first few terms of Z when v is large, (iii) bounding E[Y] when v is small, and (iv) using an asymptotic expression for Z in Equation (1). Shmueli et al.(10) used the last approach to derive the approximation for Z and the mean as:

$$E\left[Y\right]\approx\lambda^{1/v}+\frac{1}{2v}-\frac{1}{2} \tag{7}$$

Using the same approximation for Z as in Shmueli et al.(10), the variance can be approximated as:

$$Var\left[Y\right]\approx\frac{1}{v}\lambda^{1/v} \tag{8}$$

Shmueli et al(10) suggest that these approximations may not be accurate for v>1 or $\lambda^{1/v}<10$.

Despite its flexibility and attractiveness, the COM-Poisson has limitations in its usefulness as a basis for a GLM, as documented in Guikema and Coffelt((17). In particular, neither λ nor v provide a clear centering parameter. While λ is approximately the mean when v is close to one, it differs substantially from the mean for small v. Given that v would be expected to be small for overdispersed data, this would make a COM-Poisson model based on the original COM-Poisson formulation difficult to interpret and use for overdispersed data.

Guikema and Coffelt(17) proposed a re-parameterization using a new parameter μ = $\lambda^{1/v}$ to provide a clear centering parameter. This new formulation of the COM-Poisson is summarized in Equations (9) and (10) below:

$$P(Y=y)=\frac{1}{S(\mu,\nu)}\left(\frac{\mu^{y}}{y!}\right)^{\nu} \qquad (9)$$

$$S(\mu,\nu)=\sum_{n=0}^{\infty}\left(\frac{\mu^{n}}{n!}\right)^{\nu} \qquad (10)$$

By substituting $\mu=\lambda^{1/\nu}$ in equations (4), (5), and (41) of Shmueli$^{(10)}$, the mean and variance of *Y* are given in terms of the new formulation as $E\left[Y\right]=\frac{1}{\nu}\frac{\partial \log S}{\partial \log \mu}$ and $V\left[Y\right]=\frac{1}{\nu^{2}}\frac{\partial^{2}\log S}{\partial \log^{2}\mu}$ with asymptotic approximations $E\left[Y\right]\approx \mu+\frac{1}{2}\nu-\frac{1}{2}$ and $Var\left[Y\right]\approx \frac{\mu}{\nu}$ especially accurate once $\mu>10$ and $\nu\leq 1$. With this new parameterization, the integral part of μ is the mode leaving μ as a reasonable centering parameter. The substitution $\mu=\lambda^{1/\nu}$also allows *ν* to keep its role as a shape parameter. That is, if ν < 1, the variance is greater than the mean, while ν > 1 leads to underdispersion. In this paper we investigate the accuracy of the approximation $E\left[Y\right]\approx \mu+\frac{1}{2}\nu-\frac{1}{2}$ more closely.

This new formulation provides a good basis for developing a COM-Poisson GLM. The clear centering parameter provides a basis on which the centering link function can be built, allowing ease of interpretation across a wide range of values of the shape parameter. Furthermore, the shape parameter *ν* provides a basis for using a second link function to allow the amount of overdispersion, equidispersion or underdispersion to vary across measurements.

2.2 Generalized Linear Model. Guikema and Coffelt$^{(17)}$ developed a COM-Poisson GLM framework for modeling discrete count data using the reformulation of the COM-Poisson given in equations (9) and (10). This dual-link GLM framework, in which both the mean and the variance depend on covariates, is given in equations (11-14), where *Y* is the count random variable being modeled and x_i and z_i are covariates. There are *p* covariates used in the centering link function and *q* covariates used in the shape link function. The sets of parameters used in the two link functions do not need to be identical. If a single-link model is desired, the second link given by equation (14) can be removed allowing a single *v* to be estimated directly.

$$P\left(Y = y_i \,|\, \mu_i, \nu\right) = \frac{\left(\mu_i^{y_i}\right)^{\nu}}{\left(y_i!\right)^{\nu}} \cdot \left[S\left(\mu_i, \nu\right)\right]^{-1} \tag{11}$$

where:

$$S\left(\lambda_i, \nu\right) = \sum_{n=0}^{\infty} \frac{\left(\mu_i^{n}\right)^{\nu}}{\left(n!\right)^{\nu}} \tag{12}$$

$$y_i = 0,1,2,\ldots,\infty$$

$$\lambda > 0 \text{ and } \nu \geq 0$$

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i \beta \tag{13}$$

$$\ln(\nu) = \alpha_0 + \sum_{k=1}^{q} \alpha_k z_{ik} = z_i \alpha \tag{14}$$

In the remainder of the paper we will assume a single *v* rather than a varying *v* for simplicity. That is, in our simulation exercise, we employ equations (13) and (14) as a single link function. The COM-Poisson log-likelihood for a single observation with single-link function may now be written as:

$$L\left(y_i \,|\, \mu_i, \nu\right) = \nu y_i \log\left(\mu_i\right) - \nu \log\left(y_i!\right) - \log\left[S\left(\mu_i, \nu\right)\right] \tag{15}$$

Let $\eta_i = \log(\mu_i) = x_i\beta$. The log-likelihood for the entire dataset (N observations) is given as:

$$l(Y|X,\beta,\nu) = \nu\sum_{i=1}^{N} y_i\eta_i - \sum_{i=1}^{N}\nu\log(y_i!) - \sum_{i=1}^{N}\log\left[S(\eta_i,\nu)\right] \tag{16}$$

To obtain the maximum likelihood estimates (MLE) for the parameters, we suggest using unconstrained optimization to avoid convergence issues in the iterative solution method. To use unconstrained optimization, let $\zeta = \log(\nu)$. The likelihood now takes the form:

$$l(Y|X,\beta,\nu) = \exp(\zeta)\sum_{i=1}^{N} y_i\eta_i - \exp(\zeta)\sum_{i=1}^{N}\log(y_i!) - \sum_{i=1}^{N}\log\left[S(\eta_i,\exp(\zeta))\right] \tag{17}$$

To find the MLE for the parameters, we first find the partial derivatives of the log-likelihood, *l*, with respect to the coefficients and the dispersion parameter, given as:

$$\begin{aligned}\frac{\partial l}{\partial\beta_j} &= \frac{\partial l}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\\ &= \sum_{i=1}^{N}\left(y_i\nu - \frac{\partial}{\partial\eta_i}\log\left[S(\eta_i,\nu)\right]\right)x_{ij}\end{aligned} \tag{18}$$

and,

$$\frac{\partial l}{\partial\nu} = \sum_{i=1}^{N}\left(-\log(y_i!) - \frac{\partial}{\partial\nu}\log\left[S(\eta_i,\nu)\right]\right). \tag{19}$$

These equations are solved using iteratively re-weighted least squares approaches described in Nelder and Wedderburn(23), and Wood(24).

To complete the MLE, we must obtain the information matrix at the MLE of the log-likelihood to estimate the standard errors of the coefficient estimates. Sellers and Shmueli(21) derive important results implied by the likelihood equations for finding the information matrix at the MLE which will later be useful. By setting equations (20) and (21) equal to zero, we see the following:

$$\begin{aligned}\sum_{i=1}^{N} y_i \nu x_{ij} &= \sum_{i=1}^{N} x_{ij} \left\{ \frac{\partial}{\partial \eta_i} \log\left[S(\eta_i, \nu) \right] \right\} \\ &= \sum_{i=1}^{N} x_{ij} \left\{ \frac{\dfrac{n\nu(\exp(\eta_i))^n \left(\dfrac{(\exp(\eta_i))^n}{n!} \right)^{\nu-1}}{n!}}{\left(\dfrac{(\exp(\eta_i))^n}{n!} \right)^{\nu}} \right\}\end{aligned} \tag{20}$$

and,

$$\begin{aligned}\sum_{i=1}^{N} \log(y_i!) &= \sum_{i=1}^{N} -\frac{\partial}{\partial \nu} \left\{ \log\left[S(\eta_i, \nu) \right] \right\} \\ &= -\sum_{i=1}^{N} \frac{\sum_{n=0}^{\infty} \left[\left(\dfrac{(\exp(\eta_i))^n}{n!} \right)^{\nu} \log\left[\dfrac{(\exp(\eta_i))^n}{n!} \right] \right]}{\sum_{n=0}^{\infty} \left(\dfrac{(\exp(\eta_i))^n}{n!} \right)^{\nu}}\end{aligned} \tag{21}$$

These results not only permit solution of the MLE using iteratively re-weighted least squares, but also allow the information matrix (Hessian of the log-likelihood) to be solved as a function of the “true” parameters. The information matrix at the MLE of the log-likelihood, **I**, is the expected second derivative matrix of the log-likelihood:

$$I = \begin{pmatrix} I^{\beta} & I^{\beta,\nu} \\ I^{\beta,\nu} & I^{\nu} \end{pmatrix} \tag{22}$$

The parts of the information matrix follow (equations 21-23):

I$^{\beta}$: (23)

$$I^{\beta} = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{N} x_{ij} x_{ik} \left\{ \frac{\partial^2}{\partial \eta^2} \log\left[S\left(\eta_i, \nu\right) \right] \right\}$$

$$= \sum_{i=1}^{N} x_{ij} x_{ik} \left[\frac{\sum_{n=0}^{\infty} \left[\frac{\nu(\nu-1) n^2 \left(\exp\left(\eta_i\right)\right)^{2n} \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu-2}}{\left(n!\right)^2} + \frac{\nu n^2 \left(\exp\left(\eta_i\right)\right)^n \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu-1}}{n!} \right]}{\sum_{n=0}^{\infty} \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu}} \right]$$

$$-\sum_{i=1}^{N} x_{ij} x_{ik} \left[\frac{\left[\sum_{n=0}^{\infty} \left[\frac{\nu n \left(\exp\left(\eta_i\right)\right)^n \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu-1}}{n!} \right] \right]^2}{\sum_{n=0}^{\infty} \left[\left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu} \right]^2} \right]$$

$\mathbf{I}^{\beta,\nu}$: (24)

$$I^{\beta,\nu} = \frac{\partial^2 l}{\partial \nu \partial \beta_j} = \sum_{i=1}^{N} \left[y_i x_{ij} - x_{ij} \frac{\partial^2}{\partial \nu \partial \eta_i} \log\left[S\left(\eta_i, \nu\right) \right] \right]$$

$$= \sum_{i=1}^{N} x_{ij} \frac{\sum_{n=0}^{\infty} \left[\frac{n\nu \left(\exp\left(\eta_i\right)\right)^n \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu-1} \log\left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)}{n!} \right]}{\sum_{n=0}^{\infty} \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu}}$$

$$-\sum_{i=1}^{N} x_{ij} \frac{\sum_{n=0}^{\infty} \left[\frac{n\nu \left(\exp\left(\eta_i\right)\right)^n \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu-1}}{n!} \right] \sum_{n=0}^{\infty} \left[\left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu} \log\left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right) \right]}{\left(\sum_{n=0}^{\infty} \left(\frac{\left(\exp\left(\eta_i\right)\right)^n}{n!} \right)^{\nu} \right)^2}$$

$\mathbf{I}^{\nu}$: (25)

$$I^{\nu} = \frac{\partial^2 l}{\partial \nu^2} = \sum_{i=1}^{N}\left[\frac{\partial^2}{\partial \nu^2}\log\left[S(\eta_i,\nu)\right]\right]$$

$$= \sum_{i=1}^{N}\left[\frac{\sum_{n=0}^{\infty}\left[\left(\frac{(\exp(\eta_i))^n}{n!}\right)^{\nu}\log\left(\frac{(\exp(\eta_i))^n}{n!}\right)^2\right]}{\sum_{n=0}^{\infty}\left(\frac{(\exp(\eta_i))^n}{n!}\right)^{\nu}} - \frac{\left[\sum_{n=0}^{\infty}\left(\frac{(\exp(\eta_i))^n}{n!}\right)^{\nu}\log\left(\frac{(\exp(\eta_i))^n}{n!}\right)\right]^2}{\left[\sum_{n=0}^{\infty}\left(\frac{(\exp(\eta_i))^n}{n!}\right)^{\nu}\right]^2}\right]$$

The information matrix at the MLE may now be used to obtain the standard errors of the regression coefficients. Based on large sample approximation, the general large sample limit for the parameters $\theta = (\beta, \nu)$ is[24]:

$$\hat{\theta} \sim N\left(\theta, I^{-1}\right). \tag{26}$$

Due to the GLM formulation, and our use of the Poisson distribution, the standard errors follow directly from this result (e.g., $SE_i = \sqrt{I_{ii}}$), and we can find the confidence intervals for each parameter:

$$\text{CI: } \left[\hat{\theta} - t_{1-\frac{\alpha}{2}} SE < \theta < \hat{\theta} + t_{1-\frac{\alpha}{2}} SE\right]. \tag{27}$$

The GLM described above is highly flexible and readily interpreted. It can model underdispersed datasets, overdispersed datasets, and datasets that contain intermingled underdispersed and overdispersed counts (for dual-link COM-Poisson GLM models only). While we have not presented a dual-link COM-Poisson GLM, in the dual-link case the variance may be allowed to depend on the covariate values, which can be important if high (or low) values of some covariates tend to be variance-decreasing while high (or low) values of other covariates tend to be variance-increasing. The parameters have a direct link to either the mean or the variance, providing insight into the behavior and driving factors in the problem, and the mean and variance of the predicted counts are readily approximated based on the covariate values and regression parameter estimates.

3. METHODOLOGY

To test the estimation accuracy and computational burden of the MLE implementation of the COM-Poisson GLM of Guikema and Coffelt[17], we simulated a number of datasets from the COM-Poisson GLM with known regression parameters that correspond to a range of mean and variance values. We then estimated the regression parameters of the COM-Poisson GLM using the MLE implementation and the estimated parameters were then compared to the known parameter values. In this section, we present the procedural details.

3.1 Data Simulation. In order to characterize the accuracy of the parameter estimates from the COM-Poisson GLM, 100 datasets, with 1,000 observations each, were randomly generated for each of nine different scenarios corresponding to different levels of dispersion and mean. The nine scenarios include simulated datasets of underdispersed, equidispersed and overdispersed data. For each level of dispersion, three different sample means were used: high mean (~ 20.0), moderate mean (~ 5.0) and low mean (~ 0.8). Each of these 900 datasets was then used as input for the COM-Poisson GLM, and the resulting parameter estimates were compared to the known parameter values that had been used to generate the datasets.

Initially, we simulated 1,000 values of the covariates X_1 and X_2 from a uniform distribution on [0, 1]. The centering parameter μ was then generated according to Equation (13) with known (assigned) regression parameters. Note that we did not use any covariates to generate the shape parameter. Realizations from the COM-Poisson distribution are then generated using the inverse CDF method.

The regression parameter values were selected in such a way that the shape parameter *v* was always set between 0 and 1 for simulating the overdispersed datasets, above 1 for the underdispersed datasets and approximately 1 for the equidispersed datasets. The parameters that were assigned in simulating the datasets are given in the table below. Table I summarizes the characteristics of the simulation scenarios.

3.2 Testing Protocol. The coefficients of the COM-Poisson GLM were estimated using the MLE described above, implemented in R(25).

4. RESULTS

This section consists of an assessment of the performance of the Guikema and Coffelt(17) COM-Poisson GLM for the nine scenarios mentioned above. The results concerning the accuracy of the asymptotic approximation of the mean of the COM-Poisson suggested by Shmueli et al.(10) are also discussed.

4.1 Parameter Estimation Accuracy. Table II reports the fraction of simulated datasets where the 95% confidence interval does not contain the "true" parameter value for the parameter estimates. Our results show that, generally, the true coefficient values are contained within the 95% confidence interval with 95% frequency. Some instances where this is not the case are the dispersion parameter under S1 and S7, and the intercept under S3. Estimating the dispersion parameter under the MLE approach presented may be difficult. In only two cases (S2, S6) does the 95% confidence interval contain the true value of the dispersion parameter with greater than 95% frequency. While under most scenarios, the 95%

confidence interval contains the true parameter value with greater than 90% frequency, the two highest fractions of simulated datasets where the confidence interval does not contain the "true" parameter value correspond to estimation of the dispersion parameter. The true value of the dispersion parameter may be difficult to ascertain with the level of confidence assumed by the α-level selected.

Histograms of the MLE linear and dispersion parameters are plotted and compared with the true parameters in Figures 1-3. Each figure corresponds to a specific dispersion level and each subplot corresponds to a different dataset scenario. Figure 1 illustrates histograms for the overdispersed scenarios (S1-S3). While the parameter estimates generally seem to be appropriate for all of the linear parameters, β, the dispersion parameter, ν, for the high-mean overdispersed scenario (S1) is overestimated. The dispersion parameter for the moderate and low mean (S2 and S3) scenarios appear to be more well-behaved, and the ranges of the parameter distributions in the low-mean case are much wider (1.0-3.0) than the moderate and high mean cases (0.1-0.5) for each parameter estimated. While this pattern is not replicated for the under and equidispersed scenarios, the dispersion parameter distribution for all scenarios except S3 is wider than those of its attendant linear parameter distributions. For example, Figure 2 illustrates the plots for the underdispersed scenarios (S4-S6). As with the overdispersed cases, all linear parameter estimates are well-behaved. In the underdispersed scenarios, the dispersion parameters are also well-behaved, although the dispersion parameter distribution for the high-mean case (S4) seems slightly skewed to the right. Furthermore, the low-mean scenario (S6) has a wider parameter estimate distribution for each parameter when compared to its high- and moderate-mean counterparts. Figure 3 replicates this pattern for the equidispersed scenarios (S7-S9), though the magnitudes of the parameter ranges are generally smaller than those for the overdispersed scenarios (S1-S3), but slightly larger than those for the underdispersed scenarios (S4-S6).

4.2 Prediction Bias. The bias of an estimator is defined as the difference between an estimator's expected value and the true value of the parameter being estimated. If the bias is zero then the estimator is said to be unbiased. The bias of an estimator $\hat{\theta}$ is calculated as $E(\hat{\theta}) - \theta$ where θ is the true value of the parameter and the estimator $\hat{\theta}$ is a function of the observed data.

The bias of the parameters β and ν under each scenario is calculated as the difference between their average estimates from the 100 simulated datasets and the true (or assigned) value in each scenario.

The bias of the centering parameter μ is calculated as

$$E(\hat{\mu}) - \mu = (\overline{\hat{\beta}}_0 - \beta_0) + (\overline{\hat{\beta}}_1 - \beta_1)\overline{X}_1 + (\overline{\hat{\beta}}_2 - \beta_2)\overline{X}_2 \qquad (28)$$

The bias of the dispersion parameter ν is calculated as

$$E(\hat{\nu}) - \nu = (\overline{\hat{\nu}} - \nu_0) \qquad (29)$$

where $(\overline{\hat{\beta}}_i - \beta_i)$ and $(\overline{\hat{\nu}} - \nu_0)$ are the bias in the parameters and $\overline{X}_i$ is the average value of the independent variable. Figure 4 reports the parameter bias for each parameter under each scenario. The overdispersed scenarios are shown in the left panel, the underdispersed in the middle panel, and the equidispersed are shown in the right panel. For the overdispersed scenarios, the bias did not change much for the parameters, except for the intercept parameter under S3. For the underdispersed and equidispersed scenarios, the first two linear parameters and the dispersion parameter demonstrated little sensitivity to the dispersion and mean levels, while the intercept showed a marked sensitivity to these simulated conditions. As the mean decreased in these scenarios, the bias decreased for the intercept. In addition, Figure 4 suggests that the bias is largest for the overdispersed scenarios relative to the underdispersed and equidispersed scenarios.

Figures 5 and 6 report the bias in the mean estimates, given the parameter estimates. First, consider Figure 5. This plot illustrates, for each scenario, the distribution of the bias in the mean for each simulated observation in each dataset under each scenario (N=100,000). These plots suggest that the COM-Poisson GLM is biased in inferences drawn under a low-mean scenario where underdispersion is expected (S3). Furthermore, although the range of the bias distribution is largest for the large-mean equidispersed scenario (S7), the modes of the remaining bias distributions seem close to zero. Figure 6 corroborates this observation. Figure 6 illustrates the prediction accuracy under each scenario. The prediction accuracy scatterplots reflect the information encoded in Figure 5; with the exception of S2 and S3, the COM-Poisson GLM MLE estimator appears to be approximately unbiased. While the COM-Poisson GLM performs better for high and moderate mean for all three categories of dispersion, the moderate and low-mean scenarios under under- and equidispersion seem more comparable in their levels of accuracy than the same comparison in the overdispersed scenarios. For overdispersed and equidispersed datasets, the performance is worse for all low sample mean values. The COM-Poisson GLM works well for all sample mean values for the underdispersed datasets.

4.3 Accuracy of the Asymptotic Mean Approximation. The centering parameter μ is believed to adequately approximate the mean when μ >10 based on the asymptotic approximation developed by Shmueli et al.(10). However, the deviation of μ for mean values below 10 ($\mu < 10$) has not been investigated. We chose one sample from each of the nine scenarios as a basis for estimating the accuracy of the asymptotic mean approximation. First, the μ and v parameters were calculated from the estimated parameters. We examined the goodness of this approximation by simulating 100,000 random values from the COM-Poisson distribution for a given μ and v. We then plotted the mean of the simulated values against the asymptotic mean approximation, $E\left[Y\right] \approx \mu + \frac{1}{2}\nu - \frac{1}{2}$. The results showed that the asymptotic mean approximates the true mean accurately even for $10 > E\left[Y\right] > \tilde{5}$. As the sample mean value decreases below 5, the accuracy of the approximation drops. As seen in Figure 7, the asymptotic approximation holds well for all datasets with high and moderate mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for underdispersed datasets. The accuracy drops significantly for the low sample mean values for overdispersed and equidispersed datasets. There is not much difference between the asymptotic approximation of the mean. and the true mean for $E\left[Y\right] > 1\tilde{0}$. This difference begins to increase slightly at the moderate mean values, although the difference is not high. The difference can clearly be observed for the low sample mean values, particularly for the overdispersed and equidispersed datasets. This shows that one must be careful in using the asymptotic approximation for the mean of the COM-Poisson GLM to estimate future event counts for datasets characterized by low sample mean values.

4.4 Convergence of S(μ,ν). When using the MLE approach to estimate parameters for the COM-Poisson GLM, it is necessary to compute $S(\mu,v)$, an infinite sum involving the parameters μ and v, as given in Equation 8. However, for some values of μ and v, difficulties arise in achieving convergence for the sum. Minka et al.[26] examined this issue as well.

However, they focused on finding an upper bound on the relative error in the estimation of $S(\mu,v)$. The tightness of bound is not assessed, and the practical implications of whether or not the sum will converge within a practically feasible number of terms are not addressed. We address these issues directly. Table III summarizes the combinations of μ and v requiring greater than 170 terms to achieve convergence (defined as $\varepsilon = 0.0001$) in $S(\mu,v)$. These combinations are significant because calculating $S(\mu,v)$ requires computing n!, where n is the index of the term in the infinite sum. Therefore, with combinations of μ and v requiring summation of greater than 170 terms for convergence, it is necessary to calculate the factorial of numbers larger than 170. However, the built-in factorial function in R, as in most software, cannot be used for numbers greater than 170, because 170! is the largest number that can be represented as an IEEE double precision value(27). Thus, for certain combinations of μ and v, it is necessary to use an alternate (and likely slower or less precise) algorithm for computing the factorial. However, as shown in Table III, such difficulties will only arise when using data sets with a very high mean, and are therefore unlikely to be problematic in most applications. In this paper, all combinations of values of μ and v required 170 or fewer terms to reach convergence.

5. DISCUSSION

This paper shows that the COM-Poisson GLM is flexible in handling count data irrespective of the dispersion in the data. First, the true parameters lie in the 95% confidence interval for nearly all cases, except as noted above, and are generally close to the estimated value of the parameters. The confidence intervals were found to be wider for the low mean values for both the centering and shape parameters. The bias in the prediction of the parameters and the mean also does not appear to be sensitive to assumptions we have made concerning sample mean and dispersion level, except for the intercept, whose bias decreased as the sample mean decreased under the under- and equidispersed scenarios. Even at the low sample mean values, the bias is considerably less for underdispersed datasets than for overdispersed and equidispersed datasets. Despite its flexibility in handling count data with all dispersions, the COM-Poisson GLM suffers from important limitations for moderate- and low-mean overdispersed data. The Negative Binomial (Poisson-gamma) models exhibit similar behavior(28). Second, the asymptotic approximation of the mean suggested by Shmueli et al.(10) approximates the true mean adequately for $E[Y] > 5$. This value, obtained by numerical analysis of the COM-Poisson GLM, is substantially lower than the lower bound value of 10 suggested by Shmueli et al.(10). As the sample mean value decreases, the accuracy of the approximation becomes lower. The asymptotic approximation is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for underdispersed datasets. However, the accuracy drops substantially for low sample mean values for overdispersed and equidispersed datasets.

In summary, this paper has documented the performance of COM-Poisson GLM for datasets characterized by different levels of dispersion and sample mean values. These types of count data sets are of critical importance in risk assessment in areas such as infrastructure engineering, public health, traffic safety, and terrorist threat assessment. The results of this study show that the COM-Poisson GLMs can handle under-, equi- and overdispersed datasets with different mean values, although the confidence intervals are found to be wider for low sample mean values. Despite its limitations for low sample mean values for overdispersed datasets, the COM-Poisson GLM is still a flexible method for analyzing count data. The asymptotic approximation of the mean is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data, and it is also accurate for low sample mean values for underdispersed datasets. The COM-Poisson GLM is a promising, flexible regression model for count data. It provides an important advance in the tools available to risk analysts confronted with complex, real-world count data on which to base risk assessments.

6. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (grant ECCS-0725823), the U.S. Department of Energy (grant BER- FG02-08ER64644) and the Johns Hopkins Whiting School of Engineering. All opinions in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

1. Liu H, Davidson RA, Rosowsky DV, Stedinger JR. Negative Binomial Regression of Electric Power Outages in Hurricanes. Journal of Infrastructure Systems, 2005:258-67.

2. Han S-R, Guikema SD, Quiring SM. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. Risk Analysis, 2009;29(10):1143-53.

3. Han S-R, Guikema SD, Quiring SM, Lee K, Rosowsky D, Davidson RA. Estimating the spatial distribution of power outages during hurricanes in the gulf coast region. Reliability Engineering and System Safety, 2009;94(2):199-210.

4. Guikema S, Coffelt JP. Modeling count data for non-linear, complex infrastructure systems. Journal of Infrastructure Systems, 2009;15(3):172-8.

5. Maher MJ, Summersgill I. A comprehensive methodology for the fitting of predictive accident models. Accident Analysis & Prevention, 1996;28(3):281-96.

6. Lord D, Guikema SD, Geedipally S. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention, 2008;40(3):1123-34.

7. Lord D, Mannering FL. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research-Part A, 2010;44(5):291-305.

8. Anastasopolous PC, Mannering FL. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis & Prevention, 2009;41(1):153-9.

9. Conway RW, Maxwell WL. A queuing model with state dependent service rates. Journal of Industrial Engineering, 1962;12:132-6.

10. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. Applied Statistics, 2005;54:127-42.

11. Boatwright P, Borle S, Kadane JB. A model of the joint distribution of purchase quantity and timing. Journal of the American Statistical Association, 2003;98:564-72.

12. Borle S, Boatwright P, Kadane JB. The timing of bid placement and extent of multiple bidding: An empirical investigation using eBay online auctions. Statistical Science, 2006;21(2):194-205.

13. Kadane J, Shmueli G, Minka G, Borle T, Boatwright P. Conjugate analysis of the Conway Maxwell Poisson distribution. Bayesian analysis, 2006;1:363-74.

14. Ridout MS, Besbeas P. An empirical model for underdispersed count data. Statistical Modelling, 2004;4:77-89.

15. Telang R, Boatwright P, Mukhopadhyay T. A mixture model for internet search-engine visits. Journal of Marketing, 2004;41:206-14.

16. Rodrigues J, de Castro M, Cancho VG, Balakrishnan N. COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, 2009;139:3605-11.

17. Guikema SD, Coffelt JP. A flexible count data regression model for risk analysis. Risk Analysis, 2008;28(1):213-23.

18. Khan NM, Jowaheer V. A comparison of marginal and joint generalized quasi-likelihood estimating equations based on the COM-Poisson GLM: Application to car breakdowns data. International Journal of Mathematical and Statistical Sciences, 2010;2(2).

19. Lord D, Geedipally S, Guikema S. Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. Risk Analysis, 2010;30(8):1268-76.

20. Geedipally S, Lord D. Examining the crash variances estimated by the Poisson-Gamma and Conway-Maxwell-Poisson models. 90th Annual Meeting of the Transportation Research Board; Washington, DC, 2011.

21. Sellers KF, Shmueli G. A flexible regression model for count data. Annals of Applied Statistics, 2010;4:943-61.

22. Jowaheer V, Khan N. Estimating regression effects in COM Poisson generalized linear model. World Academy of Science, Engineering, and Technology, 2009;53:213-23.

23. Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society, Series A, 1972;135(Part 3):370-84.

24. Wood SN. Generalized Additive Models: An Introduction with R. Carlin BP, Chatfield C, Tanner MA, Zidek J, editors. New York: Chapman and Hall/CRC; 2006.

25. Team RDC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.

26. Minka TP, Shmueli G, Kadane J, Borle S, Boatwright P. Computing with the COM-Poisson distribution. Pittsburgh, PA: Department of Statistics, Carnegie Mellon University2003 Report No.: TR776.

27. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes: The art of scientific computing. New York: Cambridge University Press; 2007.

28. Lord D. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention, 2006;38(4):751-66.

TABLES AND FIGURES

Table I: Assigned parameters for the case study. The scenarios are labeled S1 through S9. ν_0 given on log scale.

	Overdispersed data			Underdispersed data			Equidispersed data		
	High mean S1	Moderate Mean S2	Low mean S3	High mean S4	Moderate Mean S5	Low mean S6	High mean S7	Moderate Mean S8	Low mean S9
β_0	3.0	1.3	-2.0	3.0	1.7	0.2	3.0	1.7	0.2
β_1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
β_2	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
ν_0	-0.4	-1.3	-1.3	1.0	1.0	1.2	0	0	0

Table II: Fraction of cases where "true" parameter values do not fall in confidence interval.

Dispersion	Scenario	β_0	β_1	β_2	ν
Overdispersed	Hi Mean (S1)	0.08	0.03	0.02	0.45
	Mod Mean (S2)	0.03	0.04	0.03	0.03
	Lo Mean (S3)	0.22	0.01	0.04	0.08
Underdispersed	Hi Mean (S4)	0.05	0.04	0.04	0.07
	Mod Mean (S5)	0.04	0.01	0.05	0.09
	Lo Mean (S6)	0.04	0.04	0.05	0.05
Equidispersed	Hi Mean (S7)	0.06	0.06	0.01	0.35
	Mod Mean (S8)	0.05	0.03	0.04	0.1
	Lo Mean (S9)	0.04	0.04	0.03	0.06

Table III: Combinations of λ and ν values requiring greater than 170 terms for convergence of $Z(\lambda,\nu)$ (Equation 2).

ν	M
0.05	≥ 190
0.10	≥ 110
0.15	≥ 102
0.20	≥ 98
0.25	≥ 105
0.30	≥ 110
0.35	≥ 111
0.40	≥ 112
0.45	≥ 113
0.50	≥ 117
0.55	≥ 118
0.60	≥ 120
0.65	≥ 121

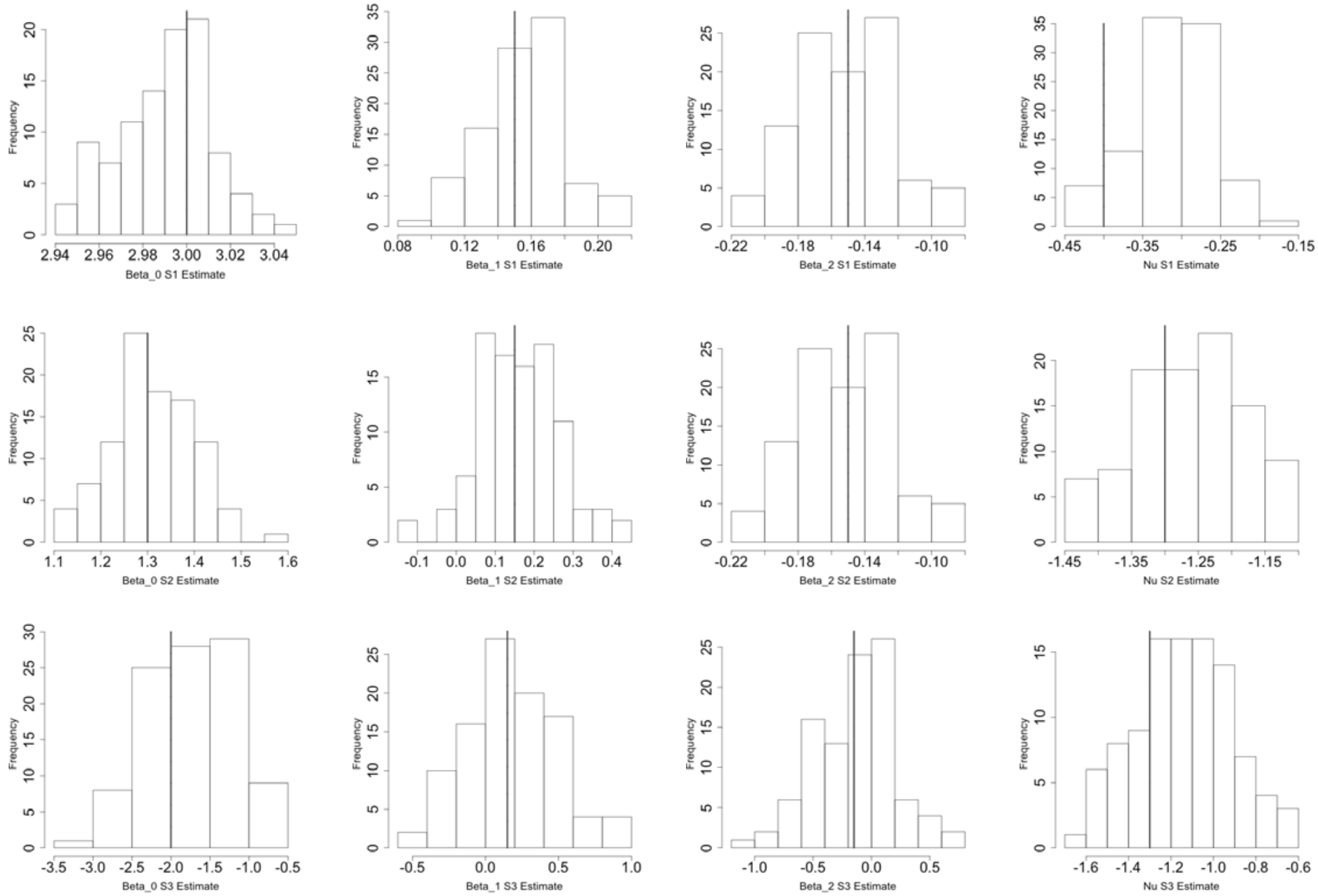


Figure 1: Histograms for parameter estimates in overdispersed scenarios. S1 (high-mean) first row, S2 (moderate mean) second row, S3 (low mean) third row. "True" parameter values indicated by vertical red lines.

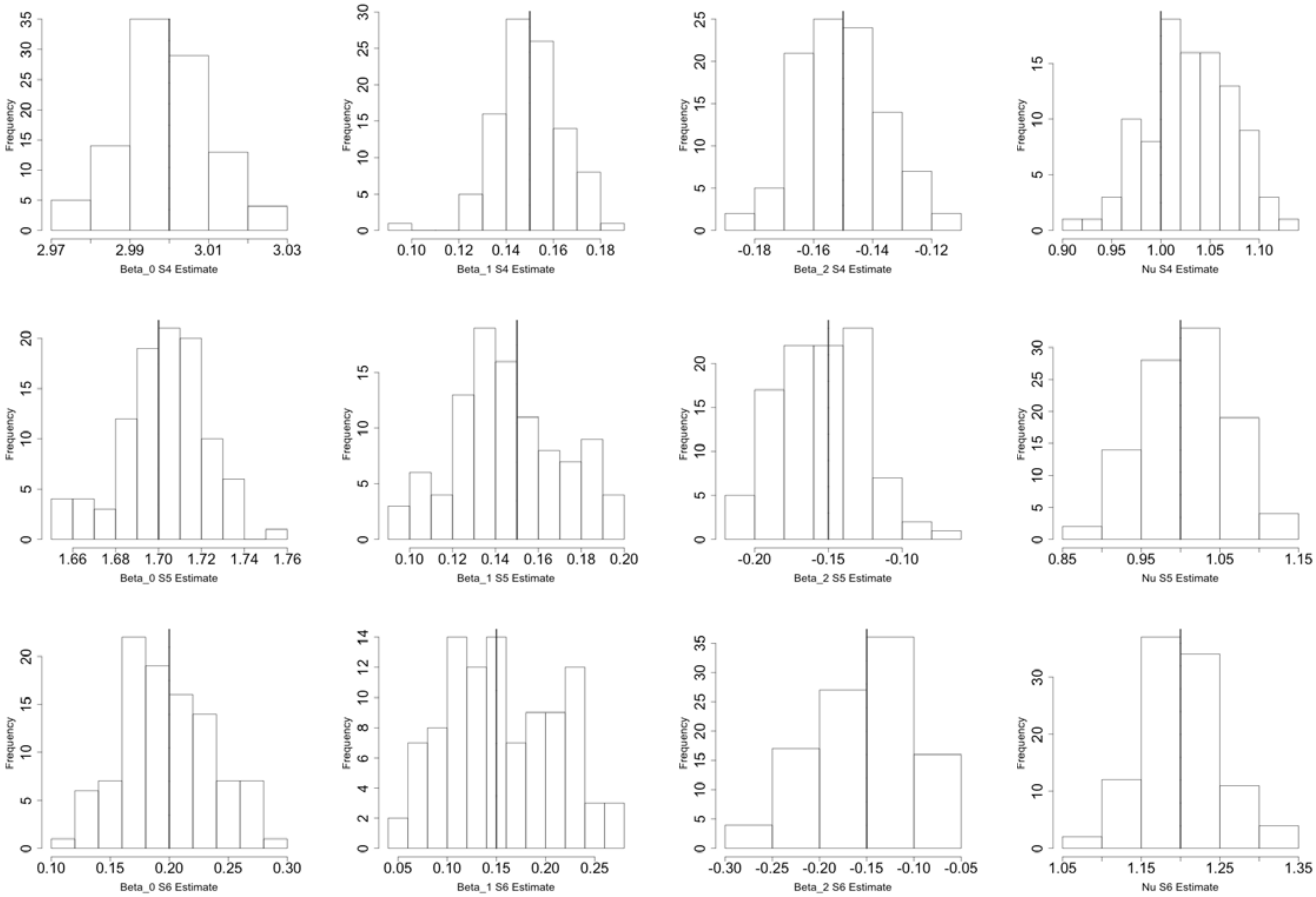


Figure 2: Histograms for parameter estimates in underdispersed scenarios. S4 (high-mean) first row, S5 (moderate mean) second row, S6 (low mean) third row. "True" parameter values indicated by vertical red lines.

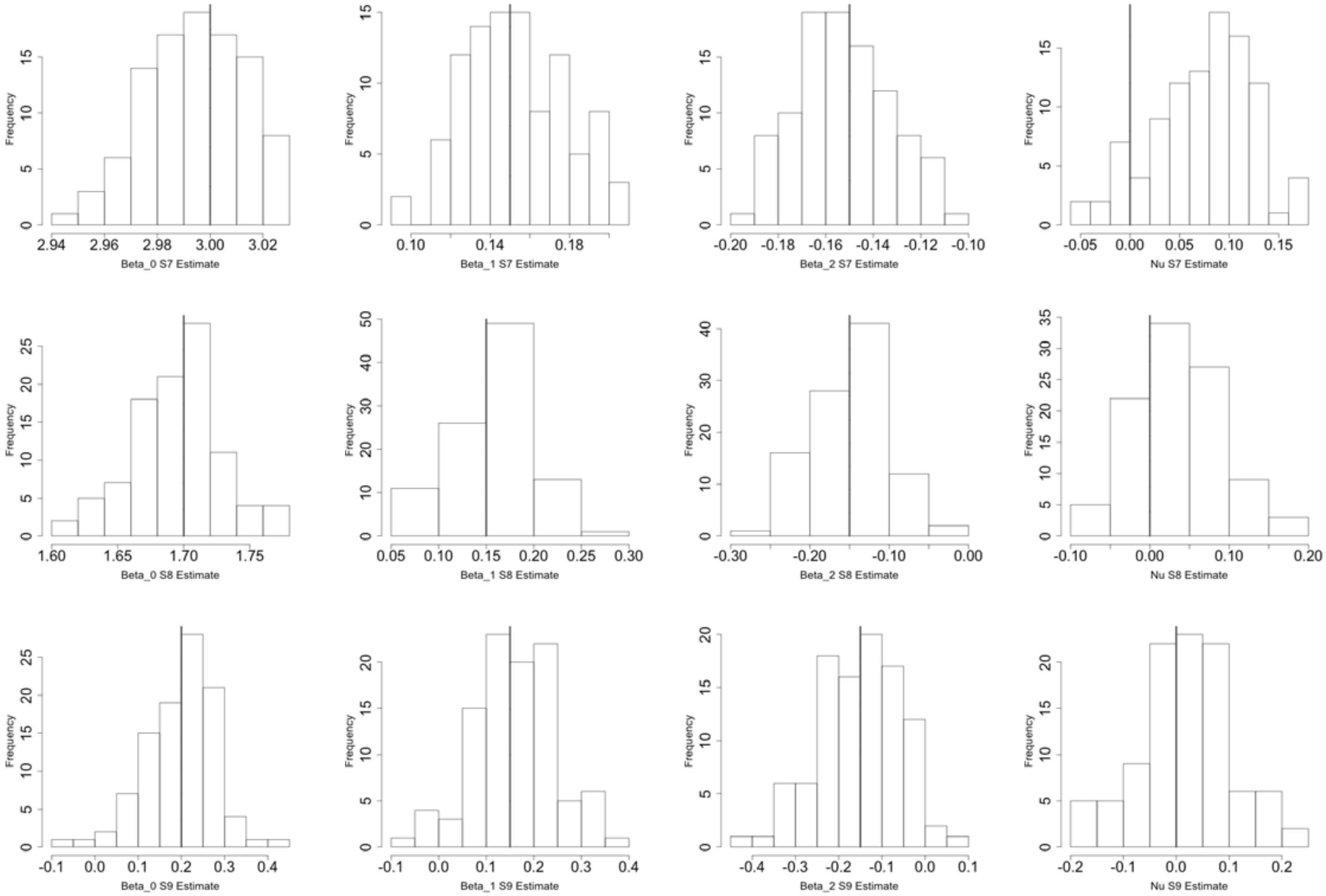
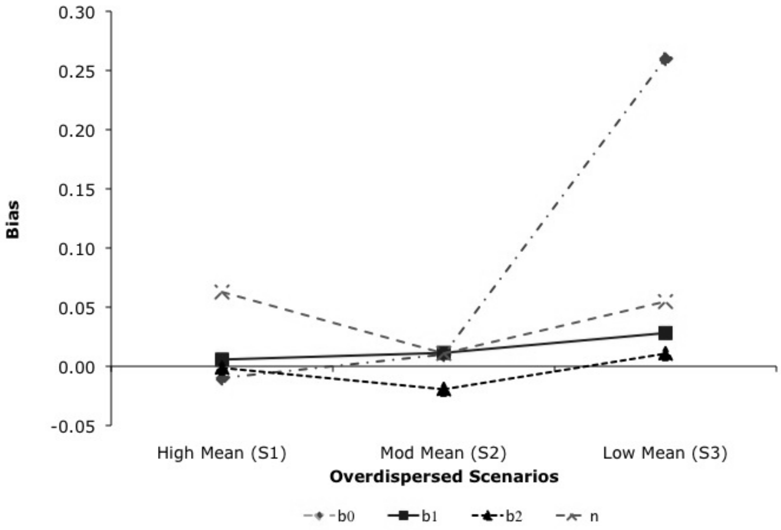
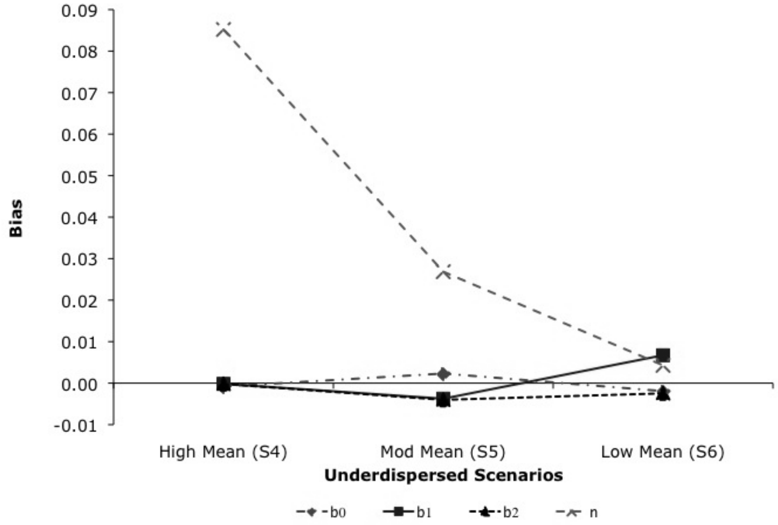


Figure 3: Histograms for parameter estimates in equidispersed scenarios. S7 (high-mean) first row, S8 (moderate mean) second row, S9 (low mean) third row. "True" parameter values indicated by vertical red lines.


0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Bias
High Mean (S1)
Mod Mean (S2)
Low Mean (S3)
Overdispersed Scenarios
b0
b1
b2
n


0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
Bias
High Mean (S4)
Mod Mean (S5)
Low Mean (S6)
Underdispersed Scenarios
b0
b1
b2
n

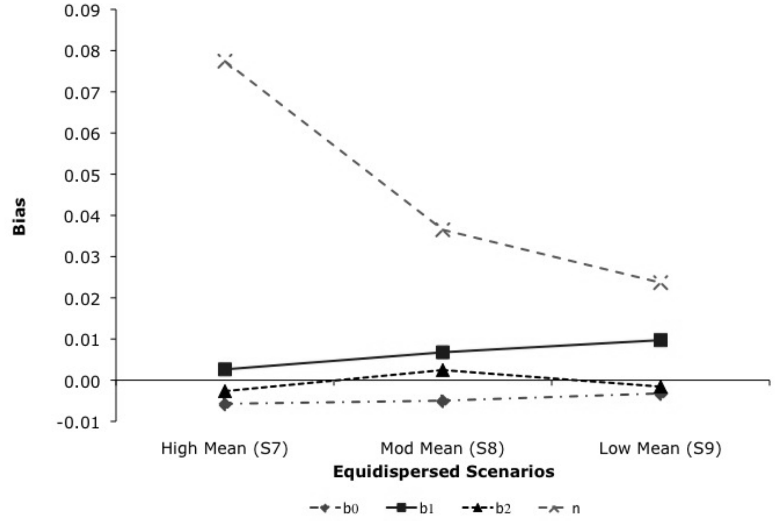

0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
Bias
High Mean (S7)
Mod Mean (S8)
Low Mean (S9)
Equidispersed Scenarios
b0
b1
b2
n

Figure 4: Prediction bias for parameter estimates under each scenario. β_0 diamonds, β_1 squares, β_2 triangles, ν indicated by "x" on line.

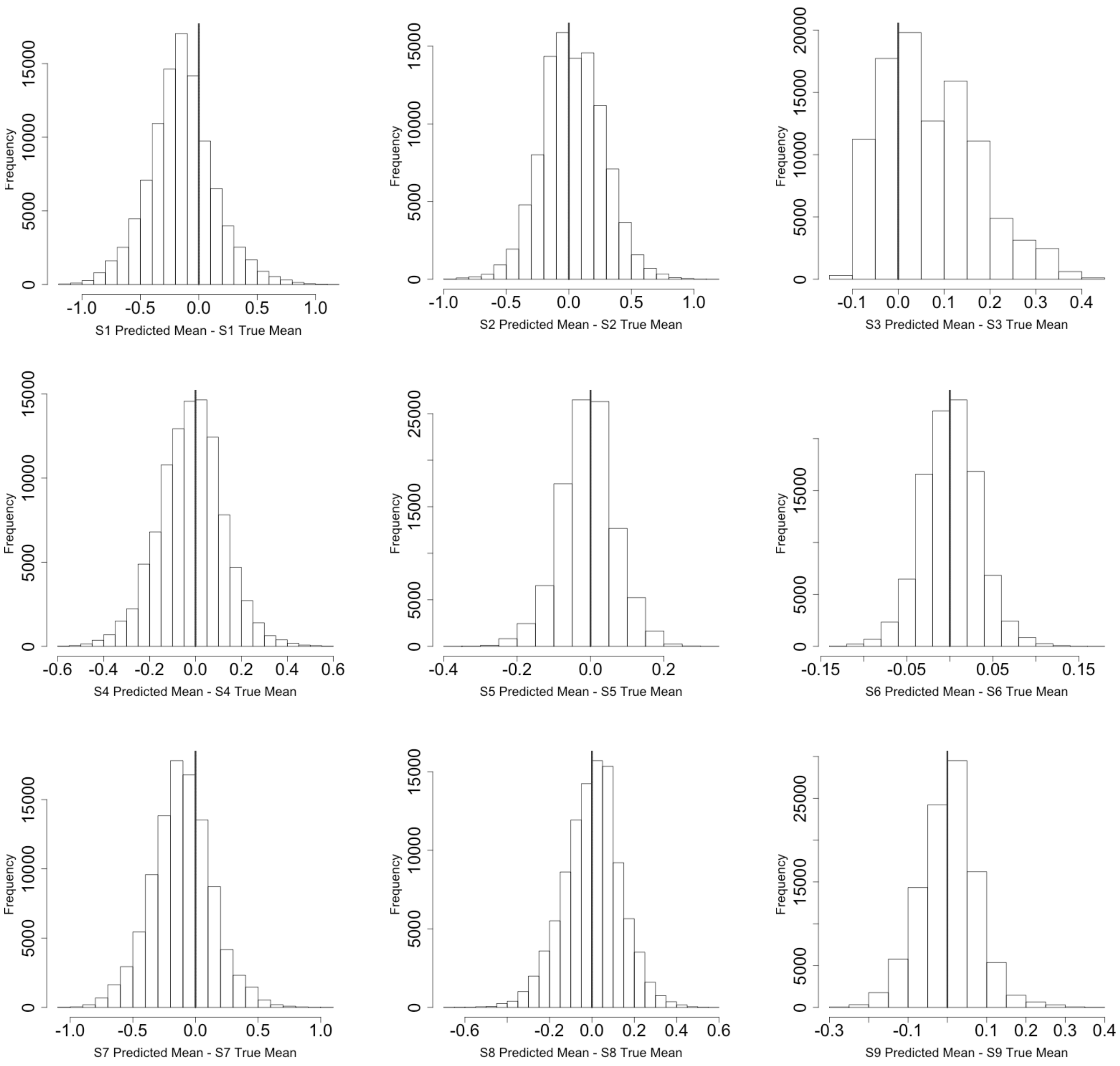


Figure 5: Prediction accuracy histograms for overdispersed (top row, S1-S3), underdispersed (middle row, S4-S6), and equidispersed (bottom row, S7-S9) datasets.

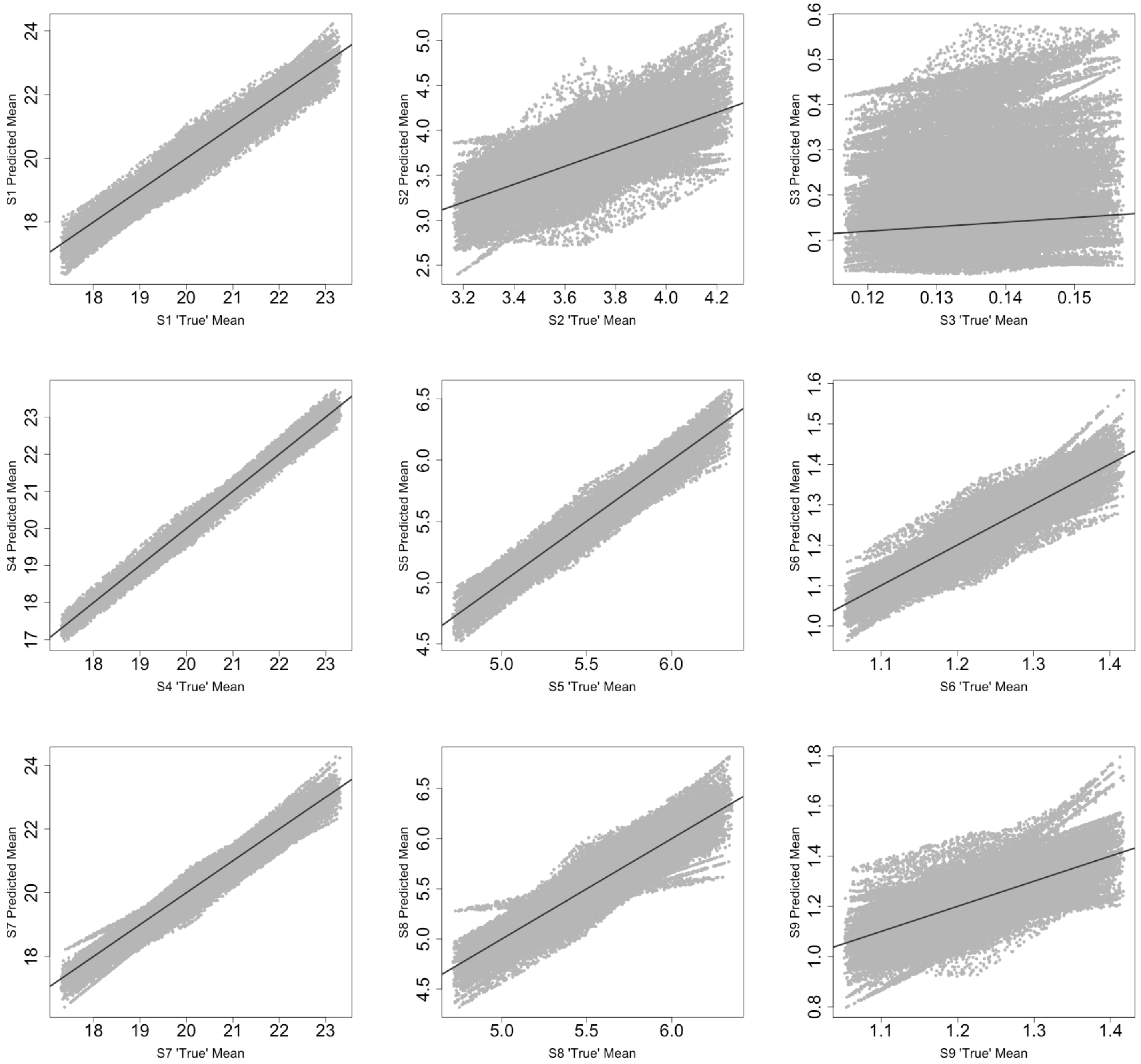


Figure 6: Prediction accuracy scatterplots for overdispersed (top row, S1-S3), underdispersed (middle row, S4-S6), and equidispersed (bottom row, S7-S9) datasets.

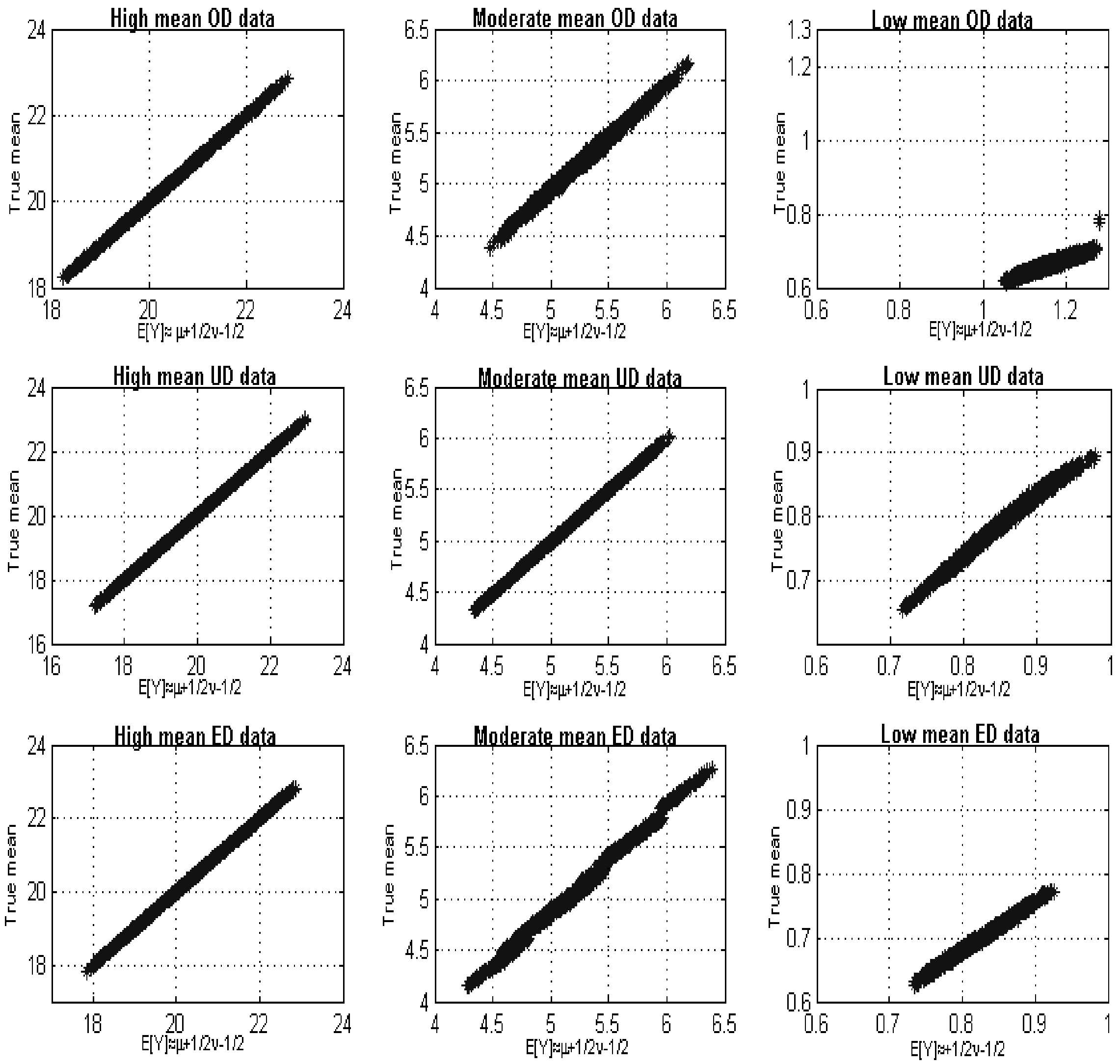


Figure 7: Mean versus asymptotic approximation for each scenario.